THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 22, Number 3/2021, pp. 283–290

A MODEL FOR PREDICTING REGIONAL DEVELOPMENT INDICATORS OF TRANSPORTATION NETWORKS USING REGIONAL INDICES

Fei LU, Yi ZHONG

Wuhan University of Technology, School of Information Engineering, Wuhan, Hubei 430070, China
Corresponding author: Yi Zhong, E-mail: zhongyi@whut.edu.cn

Abstract. The importance of transportation to economic and social development cannot be overstated. Transport equipment is transported between two nodes in different regions through the transportation network. Therefore, for the complex network of transportation networks, the proposed algorithm assumes that the role of a region as an intermediary to mediate transportation between different regions is closely related to the regional development indicators. Three indices for individual regions are proposed to quantify the degree of connectivity and importance with other regions in the transportation network through information on regional nodes of different service routes. Taking regional development indicators as dependent variables, the best model obtained by multiple linear regression on the model with permutations and combinations of explanatory variables is used to make predictions for the new data. This hypothesis is tested by using China's railroad passenger network and China's regional GDP. The results not only show that the three regional indices are good predictors of regional development indicators, but also support the idea that regions with more integrated transportation networks have better regional development indicators.

Key words: transportation networks, regional development indicators, regional indices, multiple linear regression.

1. INTRODUCTION

Transport is an important link in the production and distribution process. Networking is the basic guarantee for modern transport organisation and management. Considering that a transportation network consists of transport routes, transport stations, transport equipment and other related facilities, it is assumed that nodes denote important distribution points of transport routes, such as ports, stations, etc. The edges of the network represent the connecting channels between distribution points, such as air routes, railway lines, etc. The region is a collection containing a number of nodes and with certain boundaries between regions, such as cities, countries, etc. Early research in complex transportation networks focused on constructing the topology of networks using complex network theory to explore the statistical properties on the structure of analyzing static transportation networks. In many studies, the topology of a transportation network is usually properly represented by the concepts of space L and P [1, 2]. Wang et al [3] proposed that as long as the same train stops at any two stations, there is a connection between the nodes representing these two stations, and the transportation network constructed is defined in P-space.

A closely related field of study is devoted to the impact of transportation networks on social development, such as economic growth and population growth. Regional development indicators are used to measure the level of regional development, the most commonly used of which are gross regional product, national trade volume and regional population size. In terms of high-speed rail infrastructure, Cheng et al. [4] and Chen et al. [5] studied the impact of their new developments on the economic structure of European cities and regions. Jia et al. [6] found that the rapid development of China's regional economy was due to the large-scale construction of high-speed railways. Gao et al. [7], by analysing 25 years of economic data from China proposed the concept of inter-industry and inter-regional learning for regional economic development, stating that the development of high-speed railways increased the industrial similarity of connected pairs of neighbouring provinces. Therefore, it can be assumed that the role of a region as an intermediary in transport between different regions is closely related to the regional development indicators.

The structure of transportation network is important for the study of regional development indicators, and in network analysis the location of nodes affects the opportunities and constraints they encounter [8]. The connectivity and centrality characteristics of network nodes are key indicators of the relationship between transportation networks and regional development indicators. Among these, network connectivity describes the degree of connectivity between network nodes and reflects reachability [9]. Li et al. [10] analysed the logistics connectivity of 31 Chinese provinces over a 13-year period from 2002 to 2014, and the empirical results showed that transport connectivity had a statistically significant and positive impact on China's economic development. Meanwhile, in order to quantify the relative importance of nodes in the overall complex network, centrality indices for nodes in various types of networks have been proposed based on specific scenarios, such as degree centrality [11], closure centrality [12], and betweenness centrality [13], etc. Porta et al. [14] found that street centrality was correlated with the location of economic activities by examining the geographical distribution of three street centrality indices and their correlation with various types of economic activities in Barcelona, Spain. Ma et al. [15] used network centrality as a bridge variable to examine the coordinated coupled development between urban public transport (UPTN) and urban commercial complexes. The results show that there is a positive linear relationship between the centrality of the UPTN and the distribution of commercial complexes. However, almost all of these studies of centrality have focused on the overall centrality of the network, ignoring the extent of regional control over the transport of goods along the shortest path between pairs of regions in the network. As such we propose local centrality, an index that measures the extent to which a region acts as a transit centre for transport.

Synthesising the above analysis and research, three indices for individual regions are proposed, namely regional connectivity, local centrality and global centrality, using regions as network mediators. The degree of connectivity and importance of a region to other regions is quantified through information on nodes along different service routes. The new data is predicted by means of an optimal model obtained by multiple linear regression of models with permutations and combinations of explanatory variables, using the regional development index as the dependent variable. In this study, to demonstrate the usefulness of the three proposed regional indices in predicting regional development indicators, China's railway passenger network and China's regional GDP are used as example. The experiments show that the three proposed regional indices are effective in predicting regional development indicators.

2. PROPOSED ALGORITHM

Figure 1 shows the construction of the model for predicting regional development indicators of transportation network using special indices. The specific construction steps are as follows.

1. Construct an unweighted transportation network and six weighted transportation networks based on the fixed service schedules of transport equipment.

2. Obtain indices for each region in the network, including seven normalized and unnormalized regional connectivity, three local centralities and two global centralities.

3. Select the three explanatory variables that best explain the regional development indicators based on Pearson's correlation coefficient and a specific selection method for the explanatory variables

4. Using the regional indicators as the dependent variables, multiple linear regressions were conducted on each of the four models with the three explanatory variables arranged and combined. Based on the variance inflation factor (VIF) and the Akaike information criterion (AIC), the one model that best explains the regional indicators is obtained and the regression coefficients are used as model parameters.

5. Predictions are made using the traffic network containing the new data and the proportion of variance is used to determine how well the predictions are.

2.1 Construction of transportation network

Since the information about the important nodes of the transport routes is precisely preserved in the network structure of space P, the transportation network topology is defined in space P and an unweighted transportation network is constructed from it in the following way. First, any pair of nodes of each transport route is interconnected to form a sub-network. When edge weights are not considered, an unweighted transportation network consisting of multiple nodes and edges is obtained through overlapping the sub-networks formed by each transport route. As shown in Figure 2, the transportation network dataset is

represented as a binary network consisting of nodes and transport routes, and the binary network is projected into a one-mode network to construct a transportation network consisting of nodes.

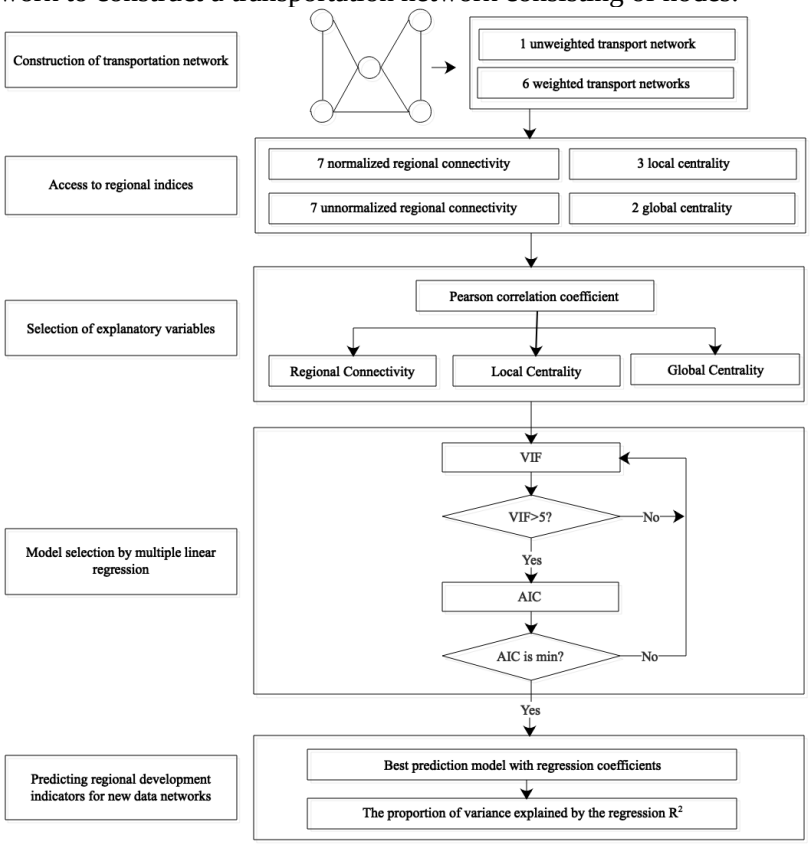

Fig. 1 Model construction diagram for predicting regional development indicators of transportation networks using special indices

Route 1	Node a->Node c->Node d
Route 2	Node a->Node b->Node e
Route 3	Node a->Node b->Node c

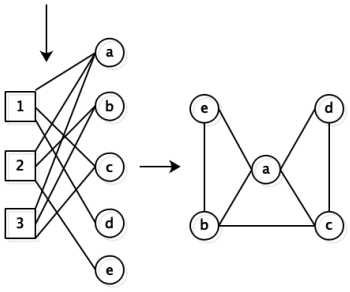

Fig. 2 The construction process of transportation network.

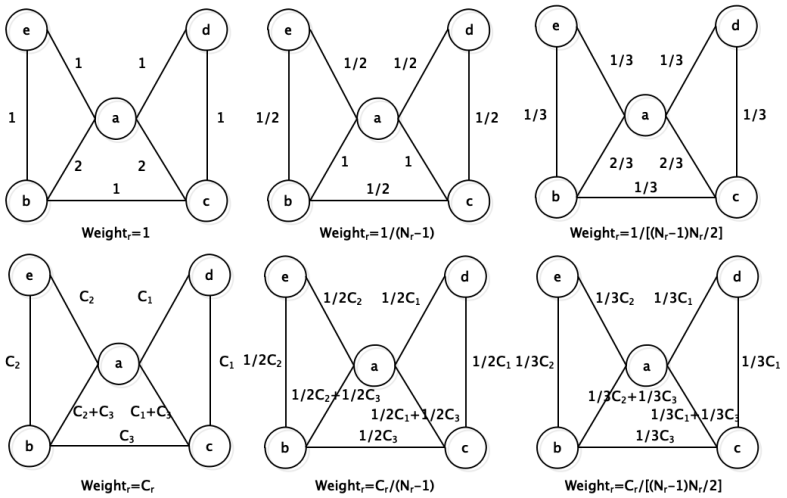

Fig. 3 Construction of weighted transportation network.

Six weighted transportation networks are constructed on the basis of the unweighted network, which is constructed in approximately the same way as the unweighted network. When the transport volume of each service route is ignored, for a transport route r containing N_r nodes, edge weights of 1, $1/(N_r-1)$, $1/[N_r(N_r-1)/2]$ are assigned to the edges of all pairs of nodes on route r, respectively. And in the case of considering the transport volume of the service route, let the total transport volume of the service route r be C_r and get the edge weight also as C_r. When the transport volume C_r of each node is equally divided by the other N_r-1 nodes on the transport route r, each edge gets the edge weight as $C_r/(N_r-1)$.

Alternatively, when considering that a transport device can transport goods in a relatively even manner between any nodes on a transport route, the edge weight $C_r / [N_r(N_r - 1)/2]$ means that C_r equally divided by all $N_r(N_r - 1)/2$ transport modes that may exist on transport route r. After obtaining the six edge weights, the edge weight between any pair of nodes is the sum of the edge weights of all the transport routes where this edge is located. Figure 3 shows the construction diagram of the weighted transportation networks.

Based on the above construction method, the edge weights between two different nodes are obtained.

$$W_{ij} = [1 - \delta(i, j)] \sum_{r \in R} Weight_r B_{ir} B_{jr} \tag{1}$$

where $\delta(i, j)$ is the Kronecker delta, R notes the set of all transport routes in the transportation network, $Weight_r$ notes the edge weight of transport route r, B_{ir} denotes whether node i is in route r, and $B_{ir} = 1$ or $B_{ir} = 0$ denotes node i is in route r and not in route r, respectively.

2.2 Access to regional special indices

Regional Connectivity: The connectivity of region a indicates how well a region is connected to other regions in the transportation network. The connectivity of region a is represented by $Rc_a = \sum_{i \in a} W_i$, where W_i is the edge weight between a node belonging to region a and a node in any other region. This definition applies to both unweighted and weighted transportation networks, thus obtaining regional connectivity for seven different edge weights. In addition, the normalized regional connectivity is also considered, which is defined as the average of the regional connectivity of all nodes in region a, denoted by $Rc_a^* = Rc_a / N_a$, where N_a is the number of nodes in region a.

Local Centrality: For local centrality, it is first necessary to define the effective path. Consider a pair of nodes i and j in different regions, and their shortest paths in the transportation network are denoted by $L_{\min}$. Since longer shortest paths represent more transit time required for transportation in network routing, only $L_{\min} = 2,3,4$ is set. When the length of the shortest path is less than or equal to $L_{\min}$ and the regions of all nodes on the shortest path except i and j are different from the regions of nodes i and j, the shortest path is an effective path. Since the regions of nodes on the effective path between nodes i and j affect the transit of goods between the regions in which the two nodes are located, they can be considered critical to the regional development indicators. When any node between i and j is in region a and there are multiple effective paths L between them, the number of effective paths of node k between i and j divided by the total number of effective paths is defined as $b_k^L = \sum_{i<j} p_k^{ij} / n^{ij}$, where L is the effective path, p_k^{ij} is the number of effective paths of node k, and n^{ij} is the number of effective paths. The sum of all b_k^L longing to the nodes of region a is defined as the local centrality of the region and denoted by $Lc_a = \sum_{k \in a} b_k^L$.

Global Centrality: The global structure information of the network is expressed in terms of global centrality, similar to the regional connectivity of the transportation network, which is defined as the sum of the betweenness centrality of the nodes in region a, denoted by $Gc_a = \sum_{k \in a} \sum_{i<j} \sigma_k^{ij} / \rho^{ij}$, where ρ^{ij} the number of shortest paths between nodes i and j, and σ_k^{ij} the number of shortest paths between i and j through node k. In addition, the same as the normalized regional connectivity, the normalized global centrality is also defined as the average of the global centrality of all nodes in region a.

2.3 Selection of explanatory variables

In order to select better explanatory variables among the obtained regional indices to be used for linear regression, Pearson correlation coefficient is used to calculate the correlation coefficient between the regional development indicators and each regional index as selection indicators. The specific method of

selecting explanatory variables is as follows. First, only regional indices with correlation coefficients greater than 0.75 can be selected. Second, the regional index with the largest correlation coefficient was selected from the seven unnormalized and normalized regional connectivity indices based on the correlation coefficient. Third, since the effective path with longer lengths contain shorter ones, try to choose local centrality when $L_{\min}$ is a smaller value. Fourth, choose the one with the larger correlation coefficient between the unnormalized and normalized global centrality.

2.4 Selecting model through multiple linear regression

Based on the selected explanatory variables, a multiple linear regression model with least squares parameter estimation was used to discover the statistical relationships with the regional development indicator. The regional development indicator for a given region a is considered as a dependent variable RDI_{a}, predicted by the combination of three explanatory variables: regional connectivity (Rc), local centrality (Lc) and global centrality (Gc). The equation obtained is

$$RDI_{\mathrm{a}} = \beta_{Rc} * Rc + \beta_{Lc} * Lc + \beta_{Gc} * Gc + \mathrm{e} \tag{2}$$

where β_{Rc}, β_{Lc}, β_{Gc} is the regression coefficient for each of the three explanatory variables.

Due to the presence of multicollinearity among the independent variables, the coefficient estimates of the multiple linear regression model may be changed erratically with small changes in the data. To obtain correlations between explanatory variables in different models to demonstrate that the model can be used for multiple linear regression in a transportation network, the variance inflation factor VIF_k was measured for each model with independent variable k [16]. When the value of VIF_k is larger, the estimated regression coefficients are also worse due to linearity. $VIF_{\max}$ indicates the maximum variance inflation factor of all independent variables in a model:

$$VIF_k = 1/(1-R_k^2) \tag{3}$$

$$VIF_{\max} = \max_{k \in M} VIF_k \tag{4}$$

where M is the set of independent variables in the model, and R_k^2 is the coefficient of determination between the kth predictor variable and the remaining predictor variables. Considering the preferred validity of multivariate linear regression in many studies with VIF less than 5 [17], the model with $VIF_{\max}$ less than 5 was chosen.

In order to select the best combination model for the explanatory variables in the multiple linear regression, the one with the smallest AIC value was given priority according to the AIC criterion. Since the ordinary least squares is used for multiple linear regression, AIC is calculated as

$$AIC = N * \ln(RSS / N) + 2K \tag{5}$$

where N is the number of observations, RSS is the sum of squared residuals, and K is the number of explanatory variables.

For each model of the multiple linear regression, the regression model is evaluated by measuring the proportion of variance explained by the regression, and the coefficient of determination is expressed as R^2:

$$R^2 = 1-(1-R^2)(N-1)/(N-K-1) \tag{6}$$

where N and K have the same meaning as in Eq. (5).

Combined with the analysis of the above studies, the best model is the one corresponding to the minimum AIC value when the $VIF_{\max}$ value is less than 5, when the coefficient of determination obtained is larger and the variance explained is better. Also, the regression coefficients of the best model are the parameters of the model expressions.

2.5 Predicting regional development indicators for new data networks

In order to evaluate the best model obtained, regional development indicator is predicted for the relevant transportation network containing the new data. The predicting steps are as follows. Firstly, the predicted regional development indicators for each region are obtained from the regression coefficients of the best model. Then total sum of squares (SST) and sum of squares due to regression (SSR) are calculated. Finally, the proportion of variance r^2 is obtained in terms of

$$r^2 = SSR / SST = \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2 \tag{7}$$

where $\hat{y}_i$ is the predicted data, y_i is the true data and $\bar{y}$ is the average of the true data. Larger values of r^2 indicate better prediction.

3. EXPERIMENTS AND RESULTS PROPOSED ALGORITHM

3.1 Experimental setting

The data on the Chinese railway transportation network for the experiments were obtained from the official website of the State Railway Administration of China (www.12306.cn). As the moment information of trains is available from this website, it is possible to obtain information on train times for 2350 in 2018 and 2613 in 2019 (The dataset can be obtained from https://github.com/trainFrequencyData/sourceData). Considering the impact of passenger traffic, only stations with Chinese railway station grades of special, first and second class were used, resulting in 723 stations in 2018 and 731 stations in 2019 for 31 provinces and cities. For the acquisition of regional GDP, the GDP of 31 provinces and cities was collected from the statistical yearbooks of the National Bureau of Statistics of China. It is important to note in particular that these data do not include information on railway trips and stations in Hong Kong, Macau and Taiwan.

For each trip on the Chinese railway network, trains can stop at railway stations in sequence and carry passengers between any two stations according to a fixed schedule. Therefore, the railway transportation network can be defined in terms of space P to obtain an unweighted railway passenger network consisting of 723 nodes versus 182,453 edges in 2018 and 731 nodes versus 183,218 edges in 2019. Since the total passenger traffic of a single train is the sum of the remaining tickets between all stations that the train stops at, the same six weighted Chinese railway passenger transportation networks are constructed.

3.2 Results

Table 1

Pearson's correlation coefficient between regional connectivity and GDP of 31 Chinese provinces and cities in 2018

Edge weight	None	1	$1/(N_r-1)$	$1/[N_r(N_r-1)/2]$	C_r	$C_r/(N_r-1)$	$C_r/[N_r(N_r-1)/2]$
r (unnormalized)	0.876	0.853	0.832	0.813	0.868	0.872	0.884
r (normalized)	0.478	0.436	0.408	0.467	0.464	0.461	0.453

Table 2

Pearson's correlation coefficient between regional centrality and GDP of 31 Chinese provinces and cities in 2018

	Local Centrality			Global Centrality	
	$L_{\min}=2$	$L_{\min}=3$	$L_{\min}=4$	unnormalized	normalized
r	0.863	0.871	0.882	0.857	0.446

Tables 1 and 2 show the Pearson correlation coefficients between each network index and the GDP of 31 Chinese provinces and cities in 2018. Since most of the correlation coefficients are greater than 0.85, many indices of China's railroad passenger transportation network are strongly correlated with gross regional product. Given the explanatory variable selection method and the results shown in Tables 1 and 2, unnormalized regional connectivity with weights equal to $C_r/[N_r(N_r-1)/2]$, local centrality with $L_{\min}=2$, and unnormalized global centrality were retained.

After obtaining the explanatory variables, multiple linear regressions are performed on four models of their permutations to explain the GDP of 31 Chinese provinces and cities in 2018, $VIF_{\max}$, AIC and R^2 are measured. The regression results are presented in Table 3. In terms of model selection, models with $VIF_{\max}$ greater than 5 are first excluded because these models generally have correlation between explanatory variables. Among the remaining models, the one with the smallest AIC value is the bivariate model consisting of regional connectivity and local centrality. It explains 86.7% of the variance in GDP across provinces and cities, yielding regression coefficients of $\beta_{Rc} = 0.547,\ CI = [0.431, 0.663]$ and $\beta_{Lc} = 0.435,\ CI = [0.319, 0.551]$, with 95% confidence intervals shown in square brackets.

Table 3

Results of the multiple linear regression with GDP as the dependent variable for 31 Chinese provinces and cities in 2018.

Model	$VIF_{\max}$	AIC	R^2
Rc + Lc	2.246	-306.24	0.867
Rc + Gc	3.675	-284.87	0.853
Lc + Gc	6.478	-266.63	0.851
Rc + Lc + Gc	13.631	-324.43	0.881

The obtained model and regression coefficients are used to predict the data on China's railroad passenger network and the GDP of 31 Chinese provinces and cities in 2019. Better results were obtained with a coefficient of determination value of 0.851, explaining 85.1% of the variance. In addition to verify that the bivariate model of regional connectivity and local centrality is the best model, multiple linear regressions are conducted for four models. The results obtained as seen in Table 4 are the same as those of the 2018 data and the best model is still the model consisting of regional connectivity and global centrality. Meanwhile, the regression coefficients of the 2019 model were obtained as $\beta_{\mathrm{Rc}} = 0.541,\ CI = [0.417, 0.665]$ and $\beta_{Lc} = 0.442,\ CI = [0.318, 0.566]$, with 95% confidence intervals shown in square brackets, and the results were also similar in magnitude to those of the 2018 model.

Table 4

Results of the multiple linear regression with GDP as the dependent variable for 31 Chinese provinces and cities in 2019.

Model	$VIF_{\max}$	AIC	R^2
Rc + Lc	2.223	-298.67	0.858
Rc + Gc	3.827	-278.78	0.841
Lc + Gc	6.726	-257.51	0.836
Rc + Lc + Gc	15.428	-307.34	0.868

3.3 Discussion

As an important network in transportation, the structure of China's railway passenger network is closely linked to the regional GDP. The position of a province or city in the railway passenger transportation network not only provides a reference basis for the train schedule design of the railroad bureau, but more importantly, affects the regional GDP. Based on the comprehensive Chinese railroad passenger transportation network dataset, the railroad passenger transportation network structure is constructed and the regional connectivity, local centrality and global centrality of this network structure are obtained. These three indices explain the regional GDP quite well, with the bivariate model consisting of regional connectivity and local centrality being the best model for predicting regional GDP in the railway passenger transportation network. Therefore, it can be concluded that the provinces and cities that are more integrated into China's railroad passenger network also have better regional economic development.

4. CONCLUSION

Transportation networks originate from the diverse designs of service routing by various freight and passenger transport companies worldwide. The structure of the transportation network physically supports

and influences the regional development indicators in the network. Based on this, the transportation network is constructed in space P and shows that the position of a region is a strong indicator for the development of the region. In particular, regional connectivity, local centrality and global centrality are proposed. They make use not only of the structure in the network, but more importantly, take into account the regional attributes of the nodes. Where regional connectivity reflects the extent to which a region is connected to other regions, local centrality reflects the extent to which a region acts as a transit centre for the transport of goods between different regions. A structural hole in a network is the absence of connections between a pair of nodes in an egocentric network. The local centrality of the transportation network quantifies the number of structural holes in the nodes of a given region, in particular the open triangle consisting of three nodes located in different regions for $L_{\min} = 2$. As the local centrality supports the structural hole theory, nodes with more structural holes can better explain the regional indicators, and therefore the strong correspondence between the local centrality and the regional development indicator suggests that structural holes occupying between nodes in other regions may yield higher regional development indicators. A multiple linear regression of the four models with a linear combination of the three explanatory variables was performed and the best model obtained was used to predict regional development indicators for the same transportation network containing the new data.

In addition, it is important to note that the model can be applied for transportation networks that contain at least three nodes belonging to different regions on a single transport route. However, for transportation networks that contain only start and end nodes, it is expected that the obtained indices will not provide a good prediction of regional development indicators. Furthermore, many transportation networks are not single networks, but rather composite networks formed by superimposing multiple transport sub-networks. Therefore, it will be a future work to analyze in depth the causal relationship between composite transportation network and regional development indicator.

REFERENCES

1. P. SEN, S. DASGUPTA, A. CHATTERJEE, P.A. SREERAM, G. MUKHERJEE, S.S. MANNA, *Small-world properties of the Indian railway network*, Physical Review E, **67**, 3, pp. 036106, 2003.
2. J. SIENKIEWICZ, J.A. HOŁYST, *Statistical analysis of 22 public transport networks in Poland*, Physical Review E, **72**, 4, pp. 046127, 2015.
3. W. WANG, J. LIU, X. JIANG, Y. WANG, *Topology properties on Chinese rail network*, Journal of Beijing Jiao Tong University, **34**, 3, pp. 148-152, 2010.
4. Y.S. CHENG, B.P. LOO, R. VICKERMAN, *High-speed rail networks, economic integration and regional specialisation in China and Europe*, Travel Behaviour and Society, **2**, 1, pp. 1-14, 2015.
5. C. L. CHEN, R. VICKERMAN, *Can transport infrastructure change regions' economic fortunes, Some evidence from Europe and China*, Regional Studies, **51**, 1, pp. 144-160, 2017.
6. S. JIA, C. ZHOU, C. QIN, *No difference in effect of high-speed rail on regional economic growth based on match effect perspective*, Transportation Research Part A: Policy and Practice, **106**, pp. 144-157, 2017.
7. J. GAO, B. JUN, A. PENTLAND, T. ZHOU, C.A. HIDALGO, *Collective learning in China's regional economic development*, arXiv:1703.01369, 2017.
8. S.P. BORGATTI, A. MEHRA, D.J. BRASS, G. LABIANCA, *Network analysis in the social sciences*, Science, **323**, pp. 892–895, 2009.
9. S. ZHANG, Y. ZHANG, *Analysis of Network Accessibility*, in: 2015 Conference on Computer Engineering and Networks, pp. 1139-1146, 2015.
10. K.X. LI, G.Q. QI, *Transport connectivity and regional development in China*, Journal of International Logistics and Trade, **14**, 2, pp. 142-155, 2016.
11. L. LÜ, D. CHEN, X.L. REN, Q.M. ZHANG, T. ZHOU, *Vital nodes identification in complex networks*, Physics Reports, **650**, pp. 1-63, 2016.
12. G. SABIDUSSI, *The centrality index of a graph*, Psychometrika, **31**, 4, pp. 581-603, 1966.
13. L.C. FREEMAN, *A set of measures of centrality based on betweenness*, Sociometry, pp. 35-41, 1977.
14. S. PORTA, V. LATORA, F. WANG, S. RUEDA, E. STRANO, S. SCELLATO, *Street centrality and the location of economic activities in Barcelona*, Urban Studies, **49**, 7, pp. 1471-1488, 2012.
15. F. MA, F. REN, K.F. YUEN, Y. GUO, C. ZHAO, D. GUO, *The spatial coupling effect between urban public transport and commercial complexes: A network centrality perspective*, Sustainable Cities and Society, **50**, pp. 101645, 2019.
16. R. A. STINE, *Graphical interpretation of variance inflation factors*, The American Statistician, **49**, 1, pp. 53-56, 1995.
17. M.O. AKINWANDE, H.G. DIKKO, A. SAMSON, *Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis*, Open Journal of Statistics, **5**, 07, pp. 754, 2015.

Received May 17, 2021